Titre : Thomas Bayes aurait-il dû croquer le fruit interdit ?

Résumé

L'objectif de ce projet est d'expérimenter différentes techniques de pré-traitement et d'évaluer la performance de généralisation qu'elles permettent pour une variété de modèles de classification. Les même modèles seront appliqués sur deux jeux de données fondamentalement différent et permetteront de déterminer quel modèle s'applique mieux dans quel cas.

En particulier, nous voulons expérimenter avec un arbre de décision hybride visant à minimiser la covariance des composantes afin de pouvoir utiliser un classifieur Bayésien naïf. Nous comparerons cette approche à la décomposition en composantes primaires. Nous étudierons aussi l'utilisation de distributions mixtes pour construire des noyaux. Cette approche est justifiée par le fait que le jeu de données de prédiction du salaire présente une combinaison d'attributs catégoriques et qu'un arbre de décision nous permettrait de nous en affranchir.

En deuxième temps, nous expérimenterons avec les réseaux de neurones convolutifs et les auto-encodeurs afin d'étudier l'échantillon MNIST. Puisque les applications de ces modèles de reconnaissance de motifs sont quasi illimitée, nous pourrons maîtriser des outils capable de traiter des problèmes plus complexes comme la reconnaissance de motifs de chaleur précurseurs aux feux de forêt sur de l'imagerie satellite.

1 Pré-traitements

En plus d'appliquer les algorithmes qui suivent sur nos données, nous allons également vérifier l'impact des pré-traitements suivants :

— décomposition en composante primaires (PCA)
— arbre de décision minimisant la covariance de sorte que l'hypothèse d'indépendance soit plus justifiée pour l'utilisation d'un Bayes naïf.
— auto-encodeur
— réseau convolutif

Nous allons également explorer la possiblité d'utiliser des auto-encoders pour traiter l'image. Une approche serait d'associer aux éléments d'une classe une permutation aléatoire d'exemplaires à reproduire de sorte que l'encodeur ne puisse se contenter d'apprendre la fonction identité, mais plutôt une représentation intermédiaire potentiellement utile.

Nous avons choisi le modèle de réseau de neurone convolutif d'une part pour sa popularité et aussi pour mettre en pratique et explorer des outils comme PyTorch. Le modèle de réseau convolutif consiste à combiner l'information locale d'une image en appliquant le même réseau de neurone sur différentes régions et de mettre en commun les résultats pour construire une représentation plus compacte. L'avantage de ce type de réseau est qu'il est invariant aux translations des caractéristiques de l'image d'entrée et qu'il nécessite peu de pré-traitement puisqu'il construit lui-même une bonne représentation de l'image.

Le pré-traitement utilisé dépendra du jeu de données.

2 Algorithmes

Le classifieur de Bayes nous intéresse beaucoup, car il offre un bon cadre pour expérimenter divers modèle de densité à priori et par conséquent d'astuce de noyau. De plus, la variante naïve est une bonne base de référence pour comparer les modèle d'apprentissage statistique.

Le modèle de Bayes se base en deux temps : on estime une densité à priori sur la classe $P[y^i = c_j]$ et ensuite on ajoute l'information sur l'exemplaire $P[y^i = c_j|x^i]$. On utilise la règle de Bayes pour

calculer la probabilité à postériori. Finalement, on choisit la classe c_j associée à la plus grande probabilité.

Un arbre de décision est un algorithme permettant de classifier un exemplaire en suivant des arcs étiquetés par des prédicats sur ses attributs. Chaque feuille est étiquetée par une classe. La topologie et les attributs sont choisis en fonction d'hyper-paramètres et d'une métrique au moment de l'entraînement. Par exemple, on pourrait chercher à maximiser le gain d'information (i.e. différence d'entropie) résultant du choix d'un attribut.

Le réseau de neurone multi-couche (i.e. perceptron multi-couche) est un modèle inspiré des connexions entre neurones et synapses dans le cerveau. Il est composé d'arcs pondérés et de noeuds possédant une fonction d'activation produisant une sortie sur une combinaison linéaire des activations d'entrées pondéré par les poids des arcs. Nous allons expérimenter les différentes architectures pour adresser le problème de classification de salaire et d'image.

Les algorithmes seront appliqués sur chaque jeu de données et seront combinés avec des prétraitements spécifiques.

3 Bases de données

3.1 Prédiction du salaire

Cette base de données est destinée à prédire si une personne gagne plus de 50000$ annuellement. Cette base de données est constituée de 14 attributs (6 dont la distribution est continue et 8 dont les valeurs sont catégorielles) et d'une sortie $y \in (> 50K, <= 50K)$. Elle contient 48842 observations, dont certaines sont manquantes, recueillies aux États-Unis en 1994. Cette base de données a l'avantage d'être déjà structurée et d'avoir des attributs bien définis.

3.2 MNIST

MNIST est une base de donnée d'image de numéros décimaux écrits à la main contenant 60000 exemplaires d'entraînement et 10000 exemplaires de test. Elle a été très populaire au début des années 2000 et est encore utilisée utilisé de nos jours pour évaluer la performance de généralisation de réseaux de neurones.